DOCUMENT RESUME

ED 313 890 EC 222 011

AUTHOR Skiba, Russell J.
TITLE Temporal Parameters in the Sampling of Behavior: The Accuracy and Generalizability of Observation. Instructional Alternatives Project. Monograph No. 10.
INSTITUTION Minnesota Univ., Minneapolis.
SPONS AGENCY Office of Special Education and Rehabilitative Services (ED), Washington, DC.
PUB DATE Apr 89
GRANT G008430054
NOTE 49p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Behavioral Science Research; *Classroom Observation Techniques; Elementary Secondary Education; Generalization; *Observation; *Research Methodology; Sampling
IDENTIFIERS *Time Sampling

ABSTRACT

This literature review addresses the technical adequacy of time sampling observation systems, which is fundamental to ensuring the generalizability of data obtained from behavioral observation. Investigations comparing partial interval, whole interval, and momentary time sampling have yielded consistent results. These suggest that duration of behavior is underestimated by whole interval sampling, overestimated by partial interval sampling, and on average accurately estimated by momentary time sampling. Momentary time sampling does not yield data suitable for the estimation of frequency. Partial interval sampling may yield consistent, but not necessarily absolutely accurate, estimates of frequency. Attempts to define the "ideal" interval for time sampling have generated inconsistent results. Investigations of momentary time sampling have yielded recommended values ranging from less than 30 seconds to 5 minutes. It appears that accuracy of momentary time sampling may be most dependent on overall frequency of observation, necessitating attention to both interval length and the length of the observational session. Little attention has been paid to issues of the representativeness of a behavioral sample drawn from a brief, fixed duration observation. The available literature suggests that massing observations within a fixed duration session may prove less generalizable than will observations spread over time. A list of 114 references is included. (Author/JDD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED313890

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

University of Minnesota

MONOGRAPH NO. 10

TEMPORAL PARAMETERS IN THE SAMPLING OF BEHAVIOR: THE ACCURACY AND GENERALIZABILITY OF OBSERVATION

Russell J. Skiba

INSTRUCTIONAL ALTERNATIVES PROJECT

April, 1989

EC 222011

BEST COPY AVAILABLE

MONOGRAPH NO. 10

TEMPORAL PARAMETERS IN THE SAMPLING OF BEHAVIOR: THE ACCURACY AND GENERALIZABILITY OF OBSERVATION

Russell J. Skiba

Instructional Alternatives Project

University of Minnesota

April, 1989

Abstract

Questions about the technical adequacy of time sampling observation systems have been given less attention than questions about the accuracy of the observer, even though technical adequacy questions are fundamental to ensuring the generalizability of the data obtained from behavioral observation. Investigations comparing partial interval, whole interval, and momentary time sampling have yielded highly consistent results. These suggest that duration of behavior is underestimated by whole interval sampling, overestimated by partial interval sampling, and on average accurately estimated by momentary time sampling. Momentary time sampling does not yield data suitable for the estimation of frequency. Partial interval sampling may yield consistent, but not necessarily absolutely accurate, estimates of frequency. Attempts to define the "ideal" interval for time-sampling have generated inconsistent results. Investigations of momentary time sampling have yielded recommended values ranging from less than 30 sec to 5 min. It appears that accuracy of momentary time sampling may be most dependent on overall frequency of observation, necessitating attention to both interval length and the length of the observational session. Little attention has been paid to issues of the representativeness of a behavioral sample drawn from a brief, fixed duration observation. The available literature suggests that massing observations within a fixed duration session may prove less generalizable than will observations spread over time. Implications for the design of time sampling observational systems are discussed.

This project was supported by Grant No. G008430054 from the U.S. Department of Education, Office of Special Education and Rehabilitative Services (OSERS). Points of view or opinions do not necessarily represent official position of OSERS.

Temporal Parameters in the Sampling of Behavior: The Accuracy and Generalizability of Observation

Behavioral observation involves the interaction of an observer with an observational system for the purpose of collecting data on persons or situations. In order for the data generated through observation to be considered accurate, it is necessary to demonstrate the technical adequacy of both the observers and the observational system. Given the numerous threats to reliability posed by the observer due to reactivity (Baum, Forehand, & Zegiob, 1979; Mercatoris & Craighead, 1974), observer expectations (Kent, O'Leary, Diament, & Dietz, 1974), or lack of observer agreement (Jones, Reid, & Patterson, 1975; Kazdin, 1977), it is not surprising that there has been a significant amount of attention paid to questions of observer adequacy. Even more fundamental, though perhaps less well attended to (Foster & Cone, 1980), is the adequacy of the measurement system itself. The most well-trained and reliable of observers, operating under the least reactive of conditions, will provide accurate, generalizable data only if the measurement system itself contains no systematic error or bias. The adequacy of the observational procedures might be said to provide the upper bounds on the gains in accuracy made through attention to issues of observer bias.

Baer (1986; Baer & Fowler, 1984) has suggested that time sampling represents the system of observation most suitable for data collection in applied settings. Although measures of rate or frequency are typically regarded as the unit of analysis most likely to reflect fundamental and important qualities in the experimental analysis of behavior (Sidman, 1960), the complexity of human behavior in natural settings may make it difficult to discriminate onsets and offsets of observed behaviors. Baer and Fowler conclude that the "loose contingencies" typical of applied settings indicate that time sampling is "not something resorted to in desperation, but instead is the measure of choice" (p. 158).

Yet the introduction of time sampling methodology raises fundamental questions concerning the accuracy and generalizability of such measures with respect to the ongoing stream of behavior. Are all methods of time sampling equally

free from systematic bias with respect to observed behavior? What is the optimal length of sampling interval that will provide both efficiency of measurement and accuracy with respect to continuous measures of observed behavior? To what extent is a sample of behavior collected over a relatively short observation period sufficient to allow inferences to a broader population of behaviors occurring over longer time periods? To the degree that such questions can be answered satisfactorily, confidence in systematic behavioral observation can be assured, and the data gathered under favorable conditions (i.e. well trained and reliable observers) can be considered generalizable to some broader universe of behavior. The purpose of the present review is to address these questions by summarizing the literature pertaining to the temporal parameters of time sampling.

Development and Types of Time Sampling Methodology

Early Child Development Investigations.

The development of time sampling methodology was a direct outgrowth of child development research projects sponsored by the National Research Council in the 1920s and 1930s. Arrington (1939) suggested that dissatisfaction with the unsystematic nature of early event tallies (Bott, Blatz, Chant, & Bott, 1928) or diary records (Berne, 1930) led to the development of methods of time sampling. The method was viewed as a means of establishing experimental control in the field setting, to compensate for decreased control over environmental conditions. Arrington (1943) notes that, "In contrast to the experimental method, . . . the observer, the method of recording, and the manner of selecting the behavior to be observed are subject to control rather than the situation in which observations are made" (p. 82).

Olson (1929) is credited with the development of the first systematic use of time sampling, in a study of the "nervous habits" of children. Fourteen 10 minute observation periods were scattered over 8 school days, the dependent measure being the number of 10 minute intervals in which at least one occurrence of the behavior

was noted. A number of child development studies (Goodenough, 1928; Parten, 1932) used this observational paradigm, which has come to be known as partial interval sampling. As investigators noted that decreases in the length of the observational interval led to increases in reliability (Olson, 1931), shorter interval lengths, as brief as 5 to 10 second, were used in subsequent investigations (Arrington, 1939; Goodenough, 1930). Wasik (1984) has suggested that the most noteworthy aspect of these early investigations was the identification of factors affecting the reliability, generality, and validity of the data. Early investigators were also concerned about the representativeness of the samples of observed behavior, the reactive effects of observation, and the effects of frequency of behavior observation on the validity of the data (Arrington, 1943).

Behavioral Assessment

Observational approaches in general, and time-sampling methodology in particular, fell into disuse in the 1940s and 50s (Parke, 1979; Wasik, 1984). The emphasis placed on observational methodology by behavior analysis approaches, however, led to a renewal of interest in the method (Bijou, Peterson, & Ault, 1968). The partial interval method first developed by early child development researchers has been widely used in applied studies of behavior (e.g., Hall, Lund, & Jackson, 1968; O'Leary, Drabman, & Kass, 1973. Ragland, Kerr, & Strain, 1981; Witt & Adams, 1980). Momentary time sampling, first introduced by Bindra and Blond (1958) as a laboratory measure of general activity has also come to be widely employed in investigations of human behavior (e.g., Bushell, Wrobel, & Michealis, 1968; Hallahan, Marshall, & Lloyd, 1981; Leach & Dolan, 1985; Risley, 1971; Tindal & Parker, 1987). Taken together, time sampling methodologies have come to represent the most widely used method of observational data collection in applied behavior analysis. Kelly (1977) reported that 20% of the studies published in the Journal of Applied Behavior Analysis between 1968 and 1975 used some form of interval recording, while 21% employed momentary time sampling.

Types of Time Sampling

In its broadest sense, time sampling[1] summarizes a stream of behavior by segmenting time into discrete intervals, within which the occurrence or non-occurrence of a single instance of behavior is noted and recorded. Powell, Martindale, and Kulp (1975) have categorized methods of time sampling into three types. In partial interval sampling (also referred to as *one-zero* sampling or *Hansen frequencies*), any instance of the target behavior during the interval is counted as an occurrence. In contrast, in whole interval sampling, the observed behavior must continue for the entire length of the interval in order to be recorded as an occurrence. Methods of interval sampling have been used in conjunction with a brief recording period (i.e., 10 seconds observe, 5 seconds record), or without such a period (i.e., 10 seconds observe, with no specific period for recording). In momentary time sampling (also known as *instantaneous* or *scan* sampling), an occurrence is defined as the presence of the target behavior at the exact moment of the beginning or end of the defined interval. At very brief durations, momentary time sampling and interval recording are equivalent; for example, a 10 second momentary time sample is probably indistinguishable from a "1 second observe, 9 seconds record" partial interval method.[2]

Comparative Accuracy of Time-Sampling Methods

There has been considerable discussion concerning the appropriate measure of time sampling in the literature of behavioral assessment, child development, and ethology. The "absolute" accuracy (Kazdin, 1982) of time sampling methodologies has been assessed by comparing the results obtained through interval recording and momentary time sampling with a continuous measure of duration or frequency. Descriptive, parametric, and non-parametric analyses of the results appear fairly consistent in documenting serious concerns regarding both the accuracy of interval recording with respect to duration, and the accuracy of momentary time sampling with respect to frequency.

Accuracy with Respect to Duration

An extensive body of literature in behavioral assessment and child development has documented the relative accuracy of momentary time sampling, as well as the general inadequacy of interval recording, for providing estimates of total duration. Powell et al. (1975) compared the results generated by partial interval recording, whole interval recording, and momentary time-sampling with a continuous measure of behavior. The in-seat behavior of a secretary over two 20 minute observation sessions was videotaped, and the recording analyzed at interval lengths ranging from 10 to 120 seconds for partial and whole interval recording, and from 10 to 600 seconds for momentary time sampling. For inter-observation intervals up to 120 seconds, momentary time sampling agreed quite closely with the continuous measure of duration, and produced unsystematic errors that tended to approach zero when averaged across sessions. In contrast, both partial and whole interval recording introduced systematic error: Partial interval sampling always overestimated absolute levels of responding, while whole interval sampling always underestimated the continuous measure. These errors increased as the interval length increased, with errors as large as 24% to 40% for 120 seconds interval lengths.

Findings concerning the relative adequacy of momentary time sampling and interval recording, as well as the direction of measurement error, have been widely replicated. Using computer simulated data, Milar and Hawkins (1976) reported that partial-interval recording with an interval length of 10 seconds overestimated a continuous record of duration by anywhere from 49% to 89%. Green and Alverson (1978) applied the three procedures to computer simulated long-duration behavior and found that only the momentary time sampling approach yielded unbiased estimates of a continuous measure of the simulated behavior. Green, McCoy, Burns, and Smith (1982) compared the inter-observer agreement of the three methods (within methods accuracy), and the absolute accuracy of each with respect to a continuous measure (between methods accuracy). In terms of between methods accuracy, partial interval sampling over estimated, whole-interval sampling

underestimated, and momentary time sampling on the average accurately estimated the continuous measure. Inter-observer agreement coefficients (within methods accuracy) were also lower on the average for partial interval sampling than for either whole interval or momentary time sampling. Murphy and Goodall (1980) found that partial interval sampling overestimated the stereotyped behavior of eight severely retarded children by as much as 56% and whole interval sampling underestimated by as much as 38%. In contrast, estimates made by momentary time sampling never varied from the continuous measure by more than 4%, and often provided exact estimates of the criterion. These and other consistent results (Dixon, 1981; Harrop & Daniels, 1986; Lentz, 1982; Powell, 1984a; Powell, Martindale, Kulp, Martindale, & Bauman, 1977) appear to provide strong support for the conclusion of Powell et al. (1977) that momentary time sampling "should be employed in investigations where duration is the response dimension of interest" (p. 332).

The only apparent exception to this finding comes from a comparative investigation of partial interval sampling and momentary time sampling conducted by Repp, Roberts, Slack, Repp, and Berkler (1976). Using simulated behavior produced over 180 minute sessions, a 15 second partial interval measure (10 second observe, 5 second record) was compared to two momentary time sample estimates of 10 minutes in length. In all instances, the partial interval measure provided results more consonant with the criterion measure. However, the criterion against which the 15 second partial interval samples and the 10 minute momentary time samples were compared was not a continuous measure, but rather a 10 second partial interval measure with no recording period. This confounding of interval length and method of sampling seriously limits the generalizability of these findings. It would in fact be rather surprising if a 10 second partial interval sample was not more highly related to 15 second partial interval sampling than to 10 minute momentary time sampling.

The error generated by partial and whole interval sampling may also be non-constant, varying across experimental conditions. Powell et al. (1977) varied the

percent of time an in-seat behavior occurred over 30 minute observation sessions, such that the behavior was scheduled to occur for 20%, 50%, or 80% of each session. Results indicated that the error produced by interval sampling varied: Overestimation caused by partial interval recording was limited to 20% when the behavior occurred 80% of the time, but ranged up to 80% when the behavior occurred only 20% of the time. These results have serious implications for partial or whole interval data collected as part of an interrupted time series design, as a change in the occurrence of a behavior across phases of an experiment could yield a change in the degree of measurement error. Milar and Hawkins (1976) demonstrated that partial interval sampling distorted the apparent magnitude of change across phases of a simulated reversal design experiment, overestimating the degree of change by as much as 72%. This non-constant rate of error may preclude the application of a statistical correction procedure to interval recording data (Murphy & Goodall, 1980). Further, under some combinations of change in frequency and duration, a reduction in the total duration of behavior could be estimated as an increase when using partial interval sampling (Murphy & Goodall, 1980; Powell, 1984a).

The temporal parameters responsible for general measurement error in the three types of time sampling are fairly well understood (Ary & Suen, 1983; Powell et al., 1977). Since partial interval recording counts only the occurrence of a behavior within an interval, any instance of the behavior, no matter how brief, will result in the entire interval being scored, thus overestimating the duration of the behavior. Conversely, when the behavior must occur during the entire interval to be scored (whole interval recording), behavior occurring for only a portion of the interval will be ignored, thus underestimating the true duration of the behavior. Momentary time sampling will record some (but not all) instances of behavior that occur for less than a complete interval, producing an estimate in which sampling errors can be expected to be randomly distributed around a mean of zero (Ary, 1984; Ary & Suen, 1983). Momentary time sampling should therefore provide relatively accurate estimates of duration, depending on the extent of sampling (Powell, 1984b).

Accuracy with Respect to Frequency

Although the majority of investigations of time sampling accuracy have regarded duration as the response dimension of interest, frequency has on occasion been chosen as the dependent variable. Powell and Rockinson (1978) investigated partial interval estimates of frequency of behavior during eleven 30 minute computer simulated sessions. Given that partial interval recording will produce an accurate estimate of frequency only when there is no more than one response observed per interval, they suggested that the proportion of intervals containing only one response represents a "validity index" for partial interval sampling. For more than two-thirds of the simulated sessions, the validity index was less than .50, suggesting that partial interval sampling would, in most cases, fail to accurately represent the dimension of frequency. Repp et al. (1976) reported similar rates of error for 10 or 15 second partial interval sampling in estimating the frequency of high rate behaviors (10 per minute). When the behavior occurred at a rate of less than one response per minute, however, 10 or 15 second partial interval recording appeared to provide accurate estimates of frequency. Harrop and Daniels (1986) reported that both 15 second partial interval recording and 15 second momentary time sampling provided inaccurate estimates of rate of computer-simulated behavior, across a range of durations and frequencies. Simulating a number of time-sampling parameters simultaneously, Rojahn and Kanoy (1985) found that a simulated observation system approximated partial interval sampling more accurately in estimated frequency of occurrence than did momentary time-sampling. Yet there were surprisingly few values of any system that were accurate to within 10% of the programmed frequency, especially when the behavior was clustered or occurred more than once per minute. Thus, although partial interval sampling appears to be more sensitive to frequency of occurrence than momentary time sampling, there are indications that neither system can be relied upon for absolutely accurate estimates of rate of responding.

Ethological Literature: Duration and Frequency

Comparisons of types of time sampling also have been reported in the ethological literature. Thiemann and Kraemer (1984) note that although specific quantitative results are not generalizable across species, general principles can be expected to hold, to a certain degree, across species. In using correlational methods more often, the ethological time sampling literature may provide a clearer picture of the applicability of the different methods for relational investigations.

In a review of observational methodology, Altmann (1974) suggested that partial interval (one-zero) scores do not reflect either absolute duration or absolute frequency accurately, but rather yield a total score that confounds frequency and duration. Total time spent engaged in a behavior would be estimated accurately only if "the behavior in question took up all of the time in each interval in which it was scored, and none of the time in the others" (p. 255). Frequency would be estimated accurately only if there was no more than one event per interval; otherwise partial interval sampling results in an overestimate of frequency of occurrence. Given that such parameters may vary across individuals, or within an individual across time, Altmann rejected the use of partial interval sampling.

Empirical investigations have, for the most part, supported Altmann's theoretical account with respect to proportion of time spent in responding. Dunbar (1976) compared one-zero (partial interval) and instantaneous sampling (momentary time sampling) in observations of the social interaction and grooming of gelada baboons. Irrespective of the length of the sample interval, instantaneous sampling provided more accurate estimates of behavior. None of the instantaneous samples, and all of the one-zero samples, were significantly different from continuous observation duration estimates. Partial interval scores always overestimated actual proportion of responding, with the degree of overestimation increasing as a function of increasing interval length. Simpson and Simpson (1977) graphically demonstrated the consequences of partial interval overestimation by observing two mutually exclusive behaviors--near mother and away from mother--of baby rhesus

monkeys. The behaviors -- occurring 58% and 42% of the time respectively -- were overestimated in all instances such that the total proportion of time spent responding always exceeded 100%. At an interval length of 480 seconds, the behaviors were estimated to occur 100% and 73% of the time, respectively. In addition, a more molecular analysis indicated that it was not possible to successfully apply a correction factor, based on Markov chain processes, to correct such error.

Results from correlational studies, however, have suggested that partial interval sampling provides better estimates of frequency of occurrence than does momentary time sampling. In an investigation of chimpanzee behavior, Leger (1977) reported correlations between partial interval scores and hourly rate that ranged from .76 and .96, with the magnitude of the correlation in direct proportion to the interval length. Rhine and his associates (Rhine & Ender, 1983; Rhine & Flanigon, 1978; Rhine & Linville, 1980) have demonstrated that one-zero sampling is almost always more highly correlated with frequency of occurrence than is momentary time sampling, and that even the moderate correlations between momentary time sampling and frequency may in fact be overestimated, due to inherent correlations between the dimensions of frequency and duration (with which momentary time sampling *is* correlated). Rhine and Linville (1980) have argued that these correlations demonstrate the superiority of partial interval sampling in estimating frequency, despite absolute error that may be caused by the method. They suggest that there is nothing inherent in absolute estimates of either frequency or duration that make them the dependent measure of choice, and that in some cases, the combined estimate of rate and duration provided by partial interval sampling may be preferable to a simple measure of either rate or frequency.

Yet the confounding of frequency and duration noted by Altmann may prove troublesome in the interpretation of one-zero scores. A re-analysis of Leger's data (Kraemer, 1979) suggests that, although partial interval sampling does represent a combination of frequency and duration, it is impossible to empirically specify the relative contribution of the two parameters. Kraemer demonstrated that two

apparently equal one-zero scores could in fact represent very different combinations of frequency and duration. Rhine and Linville's (1980) analysis suggests that the influence of frequency on one-zero scores grows stronger as interval length increases. Yet this finding may suggest that some of the apparent relationship between partial interval sampling and frequency is an artifact of measurement error. As the length of the one-zero interval increases, so does the probability of at least one response in any given interval. At large values of the interval length, there is a fair probability that there will be at least one instance of the behavior in every interval. In such instances, the rate of absolute overestimation of duration induced by partial interval sampling will be so extreme as to make statements about relationship to frequency meaningless.

Summary

If total duration is the dimension of interest, momentary time sampling appears to be the measure of choice. Partial interval sampling systematically overestimates duration by generalizing any response to the entire interval; whole interval sampling underestimates duration by ignoring all but those occurrences of behavior that take up the entire interval. The degree of error caused by interval recording appears to be dependent on the duration and response length of the observed behavior, thus increasing the possibility of a non-constant error rate not amenable to statistical correction. Since the error for interval sampling is dependent on a property of the dependent variable, either method of interval recording is likely to give spurious estimates of treatment effect when that property of the dependent variable changes, as in a single case design. In contrast, the errors caused by momentary time sampling appear to be randomly distributed. Although a single session may overestimate or underestimate the proportion of responding by chance, on average the method is highly accurate with respect to duration. For estimating the total duration of a given response during an observational session, momentary time sampling is superior to either partial or whole interval recording.

The relationship between momentary time sampling, partial interval sampling, and frequency is less clear. Since any particular occurrence of a behavior will be sampled essentially by chance in momentary time sampling, there appears to be no relationship between momentary time sampling and frequency of occurrence (Ary & Suen, 1983), except insofar as the response dimension of duration (with which momentary time sampling is correlated) is related to the dimension of frequency. Partial interval sampling appears to bear a stronger relationship with frequency, but the meaning of this relationship is open to interpretation. Although the absolute differences between partial interval results and a continuous measure of frequency are typically quite large, the two measures appear to be highly associated. Thus, interval recording may be the measure of choice if consistent (though not necessarily absolutely accurate) estimates of frequency are desired.

It has been argued that partial interval sampling represents an ideal measure of the combination of rate and duration. The non-constant nature of this confound, however, means that statements regarding behavior change based on one-zero sampling would be limited to the conclusion that either duration or frequency (or both) changed when such a change was observed, and that both dimensions may have changed when no change was observed. A more satisfactory measure if estimates of both frequency and duration are desired may be a combination momentary time-sampling/interval recording measure recommended by Powell (1984b). In that system, end of interval scoring (momentary time-sampling) is used to estimate duration, while within-interval scoring of behavior initiations (modified partial interval recording) provides an index of frequency. Tests of the combined methodology using computer-simulated data yielded high inter-observer agreement and accurate estimates of both frequency and duration, given sufficient observations. Saudargas and Lentz (1986) reported on the development of a standardized observational system, the State-Event Classroom Observation System, based on such a combined approach. They report no significant differences between the SECOS estimates and real time session durations or frequencies when used for

classroom observation of 19 student and teacher behaviors.

Length of the Time Sampling Interval

For purposes of clarity, the preceding discussion has focused on between-method comparisons of time sampling. Comparison between the methods has shown that momentary time sampling is more accurate at estimating duration than interval recording. The accuracy of either momentary time sampling or interval recording will vary, however, when within-method values of interval length or sample duration are altered. The ensuing discussion of sampling parameters will examine the apparent rationale for current practice regarding observational interval length, and the empirical findings regarding the accuracy and efficiency of such lengths.

Modal Values of the Time Sampling Interval

Time sampling can be defined as a strategy of intermittent observation; investigations of appropriate sampling interval are an attempt to determine how intermittent sampling can become and still remain methodologically sound. The ideal interval for observational assessment would maximize both accuracy and economy of observation. Certainly a prime criterion in the choice of sampling interval is the degree to which the chosen temporal values generate data that are accurate with respect to some standard, whether absolute (comparison to a continuous measure) or relative (measures of inter-observer agreement). At the same time, logistical considerations dictate some concern with measurement efficiency. Given, for instance, a longer and shorter interval of equal accuracy, a longer observation interval may be preferable in generating fewer observations (and thus less observer fatigue) per time period, or in allowing more behavior codes to be observed in a given interval.

Both Kelly (1977) and Baer (1986) have noted that the interval length most typically reported in investigations using observational assessment is a 10 second period. A number of investigators have noted, however, that there appears to be little empirical justification for that value (Powell, 1984a; Sanson-Fisher, Poole, & Dunn, 1980). A picture of typical practice was provided by a survey of 103

investigations, published in 14 major journals in psychology and education between 1968 and 1985, that had been previously considered in two quantitative syntheses of classroom behavior management strategies (Skiba & Casey, 1985; Skiba, Casey, & Center, 1985/1986). Of those investigations that used some form of time sampling, the majority used a 10 second observation interval (52.1%) or a 15 second interval (10.4%). Intervals shorter than 10 seconds (ranging from 2 to 8 seconds) were reported in 14.6% of the studies, while 22.9% used intervals ranging in length from 30 seconds to 10 minutes. Not one of the studies surveyed gave a rationale for the use of the time sampling method or observation interval chosen.

In the absence of an explicit rationale, one must assume that the widespread use of a 10 second sampling interval is based on factors other than empirical evidence. The choice may be based on considerations of observer reliability and face validity. Since the goal of measurement is to accurately summarize the "stream of behavior," absolute values of the observation interval that most closely approximate continuous measurement may appear to be most valid. Mattos (1971), for instance, suggests that a 10 second interval may fairly approximate continuous recording, while being long enough to facilitate observer reliability. Alternately, investigators may simply be following the lead of early "classic" studies that recommended or used a 10 second interval (e.g., Bijou et al., 1968; Hall et al., 1968; O'Leary, Kaufman, Kass, & Drabman, 1970). Despite the widespread acceptance of a 10 second interval as appropriate for observational assessment, empirical evaluations suggest that a 10 second interval may compromise the accuracy of partial interval recording, and the efficiency of momentary time sampling.

Partial Interval Recording

In terms of absolute accuracy with respect to measures of duration, available research suggests that the optimum interval length for partial interval recording may be considerably shorter than 10 second. Simpson and Simpson (1977) reported that the shortest partial interval sample they employed, 7.5 seconds, still overestimated a continuous duration measure. Comparing interval lengths ranging

from 5 seconds to 300 seconds with a continuous measure of time in-seat, Powell et al. (1977) found that only a 5 second interval gave accurate estimates when using either partial or whole interval recording. Other results suggest that there may be no interval length at which one can be absolutely confident in the ability of partial interval sampling to detect true duration. Murphy and Goodall (1980) found that a 2.5 second observation interval provided more accurate results than a 10 second interval when using partial interval recording, but that even the 2.5 second interval overestimated the criterion measure of percent duration by as much as 26%. For interval lengths ranging from 5 seconds to 60 seconds, Tyler (1979) concluded that "one-zero sampling always overestimates, whatever the time interval and for all types of behavior" (p. 807). Powell (1984a) compared the effects of various combinations of behavioral frequency and duration on partial interval recording with interval lengths of 5 and 20 seconds. For some combinations of frequency and duration, 5 second intervals best tracked the behavior, while the 20 second interval was superior in other cases; he noted, however, that "whatever the length of the observation interval it will only be appropriate for a fairly narrow range of behavioral values" (p. 217).

Partial interval sampling should in theory be able to provide a fairly accurate estimate of frequency given a sufficiently short interval, relative to the response length and inter-response time (IRT) of the observed behavior (Ary, 1984; Suen & Ary, 1984). It can be shown mathematically that partial interval sampling will provide an accurate estimate of frequency if the shortest response length of the observed behavior is longer than the interval size, and the smallest IRT is at least twice the length of the observation interval. Further analyses have suggested that it may be possible to use partial interval sampling for estimates of duration by estimating response length and IRT with a Poisson distribution (Suen & Ary, 1986), or with a post-hoc correction based on a *z*- distribution (Suen, 1986). Since partial interval results represent a combination of the dimensions of frequency and duration, however, such corrections would be practically useful only if the relative

contributions of frequency and duration to the partial interval scores remained stable, a possibility that has been seriously questioned (Kraemer, 1979; Powell, 1984a).

It is therefore unclear to what extent the conditions specified by Suen and Ary (1986) can be satisfied in practice. While both Repp et al. (1976) and Tyler (1979) presented results suggesting that intervals up to 15 seconds provide a satisfactory measure of frequency for low to medium rates of behavior, Harrop and Daniels (1986) and Powell (1984a) found partial interval estimates of frequency to be inaccurate at almost all combinations of interval lengths, response durations, and response frequencies. Rojahn and Kanoy (1985) suggest that it should be possible to construct tables of behavioral parameters (i.e., frequency, duration, behavioral patterning), using Monte Carlo simulation, to guide the selection of interval length for measuring frequency of occurrence. Careful inspection of their data, however, reveals an extremely idiosyncratic pattern of accuracy when several behavioral parameters are considered simultaneously. Three four-way and one five-way interaction were found to be significant, making the results extremely difficult to interpret, and perhaps of limited utility in applied settings.

In summary, there appears to be no interval length at which partial interval recording accurately represents proportion of time spent responding. Theoretically, it should be possible to establish interval lengths that ensure the accuracy of partial interval recording with respect to frequency of occurrence. Both simulated and applied data suggest, however, that the interactions among various behavioral parameters may be too complex to pre-specify an accurate interval for partial interval recording of frequency. As Powell (1984a) notes:

> In practice, the determination of an "ideal" interval would have to be done session by session, and could only be accomplished if the frequency and duration characteristics of the behavior were known. Possession of this information would obviate the need for the time sampling . . . . The available weight of the evidence does support the position that partial interval sampling constitutes an inadequate measurement technique. (pp. 217-218)

Momentary Time Sampling

Early investigations that varied the length of time between momentary time samples (inter-observation interval) suggested that relatively long values could provide satisfactory estimates of duration. To validate the use of a 5 minute time sampling period, Bushell et al. (1968) compared their results with observations conducted simultaneously using momentary time sampling every 15 seconds, across three experimental phases. The plotted results appear to be quite similar for all phases, in terms of both the absolute level, and the trend of the data. Kubany and Sloggett (1973) recommended a variable-interval 4 minutes (VI 4 min) schedule of momentary time sampling for teacher-conducted observations, and compared results from that measure with a fixed-interval 15 seconds (FI 15 sec) schedule of momentary time sampling. Although particular data points from the VI 4 min schedule overestimated or underestimated the FI 15 sec observations by as much as 16%, in general the level, trend, and pattern of the two data displays appeared consonant.

Subsequent observations have suggested, however, that while relatively long values of momentary time sampling are, on average, remarkably accurate with respect to continuous measures of duration, chances of measurement error for any given session will increase as the inter-observation interval increases. Powell et al. (1975) compared values of inter-observation interval for values ranging from 10 to 600 seconds with a continuous measure of percent inseat behavior. Aggregated across sessions, estimates of proportion of behavior remained within 2% of the criterion for inter-observation intervals up to 400 seconds. For any particular session, momentary time sampling agreed quite closely with continuous measures at intervals up to 120 seconds; beyond that value, however, error increased considerably, with discrepancies as large as 74% occurring at an inter-observation interval of 600 seconds. These results have been fairly well replicated. A number of investigations (Brulle & Repp, 1984; Powell et al., 1977; Test & Heward, 1984) have reported that, although the error for any given session will increase as the inter-observation interval is lengthened, momentary time sampling will yield a very small

degree of error when averaged over sessions.

The "ideal" inter-observation interval for momentary time sampling has varied considerably across investigations. There is extensive empirical support for Brulle and Repp's (1984) conclusion that 10 second momentary time sampling is "remarkably accurate" with respect to continuous measures of duration (Green et al., 1982; Leger, 1977; McDowell, 1973; Murphy & Goodall, 1980; Powell et al., 1975, 1977), but there is less agreement concerning the appropriateness of longer inter-observation intervals. A number of investigations (Dixon, 1981; Powell et al., 1975; Tyler, 1979) have suggested that sampling intervals up to 120 seconds are satisfactory. A follow-up investigation conducted by Powell et al. (1977) indicated, however, that beyond 60 seconds, duration could be overestimated or underestimated in any particular session by 20% or more. Others suggest that the upper limit for accuracy may be even lower than 60 seconds. Observing the classroom behavior of a 10-year-old mildly handicapped child, Brulle and Repp (1984) found that samples with inter-observation intervals of 30 seconds or less generated estimates of proportion of responding that were within 3 percentage points of the continuous measure in 85% of the observations. At values of 60 seconds, 120 seconds, or 240 seconds, however, the degree of error was higher, leading them to conclude that 30 seconds was the highest momentary time sampling value that could guarantee accuracy. Similarly, Test and Heward (1984) reported that successive 30 minute observation sessions using 60 second momentary time sampling, lagged in their start point by 5 seconds, could differ as much as 18% in their estimates of a continuous measure, suggesting that data collected using that sampling interval be "evaluated with caution."

Inspection of the time-sampling data base yields a fairly consistent confound of interval length and observation session duration, however, that may contaminate current findings. With few exceptions (i.e., Mansell, 1985; Skiba, 1987), investigations of the temporal parameters of time sampling have sampled from fixed duration (usually 30 minutes) observation sessions. Investigating interval length

within a single brief value of the observation session length confounds inter-observation interval length and frequency of observation: As the length of the inter-observation interval is increased in a fixed duration session, frequency of observation decreases. A 30-minute observation period will provide a suitable number of observations at interval values of 30 seconds (120 observations), but will generate, for instance, only 6 observations when the inter-observation interval is set to 5 minutes. Walker and Lev (1953) have shown that accuracy in estimating the parameters of a finite population will improve as the number of observations in a sample (n) increases relative to the population (N) Thus, the apparent improvement in accuracy for shorter observation intervals may be due primarily to increased chances of approximating the population of behavior with a greater frequency of sampling. One would expect this relationship to be especially apparent for momentary time sampling, wherein sampling errors appear to be random (i.e., normally distributed).

The dependence of appropriate inter-observation interval length upon frequency of observation may preclude the unconditional specification of the "ideal" length. The apparent increase in accuracy that results when momentary time sampling sessions are aggregated may be due in large part to increases in reliability caused by increasing frequency of observation. Rowley (1978) demonstrated that the increase in reliability that occurs as the number of observations is increased is entirely predictable mathematically, and is roughly analogous to increases in reliability obtained by increasing test length. The inter-observation interval necessary to ensure reliability is therefore dependent upon session length; as session length decreases, interval length must likewise decrease in order to ensure some minimum frequency of observation.

Preliminary data appear to support the suggestion (Foster & Cone, 1986) that some absolute number of observations (i.e., greater than 15) may be required to achieve accuracy in momentary time sampling. In an evaluation of the suitability of time-sampling for service-evaluation research, Mansell (1985) found that for 1 hour

observation sessions, momentary time-samples of 30 seconds between observations demonstrated acceptable accuracy for relatively frequent behaviors, while 10 second momentary time sampling was required for behavior occurring for a small proportion of the observation session. When the observation was extended to 8 hours, however, intervals between observations of up to 5 minutes appeared to guarantee a fair degree of accuracy for all but the most infrequent behaviors. Similarly, Skiba (1987) investigated the effect of increasing the frequency of momentary time-sampling by observing for an entire school day. For 17 of the 19 observed behaviors, momentary time samples separated by 5 minute intervals were accurate to within 5% in 85% of the cases; a 10 minute inter-observation interval proved accurate for 12 of the 19 responses. Again, these data suggest that the accuracy of observation depends to a great degree upon the overall frequency of observation. Increasing the number of [redacted] samples, either by decreasing interval length or increasing session length, should yield a concomitant improvement in accuracy.

Theoretical treatments have suggested that the appropriate inter-observation interval for momentary time sampling also will be a function of some response dimension of the observed behavior (Haynes, 1978). Ary and Suen (1983) have demonstrated that momentary time sampling will always give an extremely accurate estimate of both frequency and duration if the inter-observation interval is shorter than both the shortest possible response length and the shortest possible IRT (inter-response time). Correction factors thus have been proposed that would allow the estimation of duration, frequency, and response length from momentary time sampling (Griffin & Adams, 1983; Suen & Ary, 1986). It seems reasonable to assume that specific response dimensions will have an influence on the appropriateness of the inter-observation interval, and computer simulation (Harrop & Daniels, 1985) has indeed suggested that the duration and frequency of behavior will influence the accuracy of momentary time sampling. Data from empirical investigations, however (Green et al., 1982; Leger, 1977; McDowell, 1973; Murphy &

Goodall, 1980; Powell et al., 1977; Sanson-Fisher et al., 1980; Tyler, 1977), have yet to reveal any striking differences in the accuracy of momentary time sampling as a function of the type of behavior observed, even when response length, frequency of occurrence, or total duration vary widely. While behavioral characteristics should, in theory, influence the choice of temporal parameters in time sampling, it may be that the interaction of frequency, duration, and response distribution is so complex as to make prediction according to some parameter of behavior problematic in applied settings.

Summary

Despite continuing recommendations for the use of 10 second partial interval sampling (Bass, 1987), there appears to be solid and consistent evidence that partial interval lengths greater than 5 seconds will result in considerable overestimation of total duration. In fact, available data suggest that there is no value for partial interval sampling that will produce consistently accurate estimates of proportion of time spent in responding. Although partial interval sampling will always provide an accurate estimate of frequency if certain conditions regarding response length and IRT are met (Suen & Ary, 1984), the complexity of actual behavior may make those conditions rare, or highly variable, in practice.

In general, momentary time sampling has been shown to provide accurate estimates of a continuous measure of behavior (expressed in terms of duration) at inter-observation interval lengths up to 30 seconds. Beyond that value, reports are contradictory. Yet the typically brief observational sessions employed in most parametric investigations has confounded interval length and frequency of observation, leaving the question of "ideal" interval length unresolved. While behavioral characteristics should in theory influence the choice of temporal parameters in time sampling, the extreme complexity of their interaction may limit the usefulness of behavioral parameters in specifying interval length.

It is becoming apparent that one of the most important parameters determining the accuracy of momentary time sampling is frequency of observation

over time (Mansell, 1985; Powell, 1984b; Skiba, 1987). Just as some minimum number of randomly selected subjects are necessary in order to justify inference to a broader population of subjects when conducting group comparison research (Cochran, 1977), a minimum number of observations that are randomly distributed (i.e., containing no systematic source of bias) with respect to the target behavior may be necessary in order to ensure generalizability to some broader universe of behavior. The decision concerning interval length thus may be a choice between intensive observation for short periods of time, or more intermittent observation occurring over extended periods of time.

Questions regarding frequency of observation over time lead inevitably to consideration of the issue of observation session length. An intensive short-duration observation session will provide results that are consonant with an intermittent long-duration session only to the extent that the patterning of behavior remains relatively consistent from the shorter to the longer session. Such questions might be termed the representativeness of the observational sample, and the data in this area raise serious concerns regarding the modal temporal parameters for behavioral observation.

Representativeness of the Observational Sample

It has been widely noted (Goldfried & Kent, 1972; Jones et al., 1975; Nelson, 1985) that behavioral assessment can be distinguished from traditional test-based assessment in terms of the level of inference drawn from the data. While more traditional approaches have tended to regard observed behavior as a sign of an underlying personality construct, behavioral assessment tends to regard behavior as a sample that is more or less likely to occur in similar stimulus situations (Goldfried, 1977; Mischel, 1968; Nelson, 1985). The data from behavioral observation thus require "little, if any, inference to unobservable or inferred constructs" (Jones et al., 1975).

Although there may be little interest in unobservable constructs in

interpreting data from behavioral observation, that is not to say there is no interest in making inferences to unobserved behavior. Foster and Cone (1986) suggest that the goal of behavioral observation is the generation of data that are "unbiased representations" of the behavior of interest, and that questions of how much and under what conditions behavior is to be observed are therefore of prime importance. Goldfried (1977) has termed such considerations sample assumptions, referring to the degree to which one assumes that the observed sample of behavior is representative of a broader population of events (that could presumably be elicited by increasing the size of the behavioral sample).

Early investigations utilizing time sampling methodologies paid considerable attention to issues of sample representativeness. Observations were rotated through the school day (Goodenough, 1928), or start times randomly chosen across a number of days (Olson, 1929), in order to ensure that the obtained sample of behavior provided a valid picture of some larger universe of responding. Goodenough (1930) noted that the demands of reliability and validity need to be carefully weighed in choosing the temporal parameters of observation. Although conducting observations during the same time period on a daily basis increased the reliability of observation, it tended to decrease the validity of the data with respect to broader samples of children's behavior. In her review of time sampling literature, Arrington (1943) concluded that the difficulty of obtaining an adequate sample was so complex that, short of extremely long observation periods, only the careful selection of samples taken from a variety of stimulus situations could ensure representativeness.

Unfortunately, more recent investigations using behavioral observation appear to have paid scant attention to issues of sample representativeness. In their 1973 review, Johnson and Bolstad reported that most research investigations in the behavior analysis literature tended to sample from "only one narrowly circumscribed situation with no evidence that the observed behavior was representative of the subject's action in other stimulus situations" (p. 50). Similar

conclusions in subsequent reviews (Foster & Cone, 1986; Goldfried, 1977; Lomax & Cooley, 1979; Wildman & Erickson, 1977) suggest there has been no increase in attention to issues of sample representativeness; a survey of recent research yields consistent findings. Among the 103 investigations cited above, only 7 investigations attempted to justify their selection of the sample of behavior observed. Only two studies (Drabman & Spitalnik, 1973; Leach & Dolan, 1985) scheduled observations so as to ensure representativeness. None of the surveyed investigations included an explicit discussion of issues of sample representativeness. The average reported value for observation session duration was 35.2 minutes with a mode of 30 minutes. Given that there appears to be no explicit justification for the use of this modal value, it seems likely that "session duration [is chosen] either arbitrarily or by considering the issue of observer fatigue" (Foster & Cone, 1986, p. 271).

Empirical Investigation of Observational Representativeness

Johnson and Bolstad (1973) commented that lack of attention to the issue of representativeness raises serious questions concerning the validity and generalizability of behavioral observation. In reviewing the validity of measures of classroom observation, Hoge (1985) reported that the the majority of reviewed studies indicated poor construct, treatment, and criterion-related validity for observational measures. The poor criterion validity of behavioral observation with respect to teacher ratings may be in part a problem of observational schedule. A brief observation session may fail to sample the high-intensity, low-frequency behaviors that appear to be highly influential in teachers' ratings (Schachar, Sandberg, & Rutter, 1986; Skiba & O'Sullivan, 1987).

Massed vs. spaced observations. The issue of observational sampling might be conceptualized as a choice between the use of massed versus spaced observations. The available literature suggests that the typical practice of massing observations in a relatively brief session may be less than optimal for representing behavior occurring over longer periods. Thomson, Holmberg, and Baer (1974) explored various methods of observation during a 64 minute session divided into 16 four

minute segments of observation. Three different sampling strategies were compared to data obtained using momentary time sampling with an interval of 10 seconds over the entire 64 minutes. In general, an intermittent sampling strategy that ensured that observations were evenly distributed across the entire session (i.e., every 4th observation segment) resulted in considerably less measurement error than did strategies that massed the same amount of observation in a contiguous time period (i.e, 16 continuous minutes of observation). Such a finding suggests that an observation period of 30 minutes scheduled contiguously could risk considerable error in estimating behavior across an entire day.

Skiba (1987) compared estimates drawn from both massed and spaced observational strategies to data collected over the course of an entire school day using 10 second momentary time sampling. Samples of 30 second, 60 second, 120 second, 300 second and 600 second intermittent classroom observation were drawn from the entire day's data using a computer simulation; similar samples were drawn to simulate massed observation in reading (30 minutes), mathematics (30 minutes), and observation begun at a random point in time (30 minutes and 60 minutes). For the majority of the 19 behaviors observed, the data from 30 or 60 minute periods of contiguous 10 second intervals proved to be a poor predictor of behavior over the entire day. In all cases, massing observations proved to be less accurate than spacing the same number of observations equally over the course of the school day.

Generalizability studies. Psychometric investigations of reliability and generalizability have yielded similar conclusions concerning the length and number of observational periods. Applying the Spearman-Brown formula to observations of 30 teachers collected over 50 minute observation periods, Rowley (1978) reported that reliability increased as the length of the observational period increased from 10 to 50 minutes. When both length and number of observations were considered simultaneously, greater reliability was achieved by the use of a large number of shorter, independent observation periods; reliability increased from .357 for one 50 minute observation period, to .516 using five 10 minute

observation periods. In a generalizability study of measures of classroom observation, Tobin and Capie (1981) reported similar findings. Generalizability coefficients for measures of individual pupil engagement increased from .12 for one observation within one lesson, to .47 for 30 observations within one lesson, to .73 for 5 observations within each of 6 lessons. Again, these results suggest the superiority of spaced, rather than massed, observation periods.

The distribution of behavior over time. In using a massed behavior sample to make inferences about a subject's typical behavior, one is assuming that the temporal distribution of the behavior during the sampling session is representative of the distribution of that subject's behavior during the time to which one is interested in making inferences. In general, behaviors that are evenly distributed over time will more likely justify such assumptions. In particular, observations massed within a brief session risk considerable error in estimating behaviors that (a) occur at a low, non-constant rate, or (b) differ significantly in their distribution across different time periods. Estep, Johnston, and Gordon (1981) reported that randomly scheduled observation sessions of up to 2 hours a day were still insufficient for accurately estimating some low frequency behaviors. In measuring student attention-to-task with a 30 second momentary time sampling procedure, Karweit and Slavin (1982) found that percent on-task was markedly higher during the first 10 minute of an instructional period than it was for the next 10 or 20 minutes. They suggested that, since time-on-task appeared to be unevenly distributed across instructional periods, inconsistencies in research estimating the relationship between academic engaged time and achievement may be due in part to measurement error caused by sampling different portions of the school day. Erlich and Shavelson (1978) found low generalizability coefficients for some teaching behaviors even when observed across 10 occasions, and suggested that some behaviors that are inherently unstable over time may never show adequate generalizability.

The distribution of behavior over time may be most strongly predicted by factors related to the situation-specificity of behavior (Bem, 1972; Mischel, 1968,

1973). The appropriateness of massing time samples into a contiguous limited duration sample may be dependent upon the degree to which the contingencies maintaining a given behavior are consistent across situations. If the contingencies for reinforcement are relatively consistent across situations, the behavior might be expected to be evenly distributed across time, and thus amenable to brief duration sampling. As contingencies maintaining a response become less consistent across situations, the ability of limited duration observation sessions to provide a generalizable estimate of behavior will decrease. Kazdin (1979) has noted that situation-specificity is in fact a "two-edged sword" that serves as a limiting condition for behavioral assessment as well as traditional assessment.

Summary and Conclusions

The complexity of human behavior and interaction provides a profound challenge for those seeking to base descriptions of behavior upon discrete periods of observation. In the face of behavioral and situational diversity, it is perhaps more remarkable that at least one strategy has been found to be relatively unbiased than it is that a number of strategies have been shown to be inaccurate. In at least some areas, as in the comparison of momentary time sampling and partial interval recording, the literature has achieved a striking consistency.

Yet it appears that current practice in behavioral observation is dictated, not so much by empirically derived parameters, as by assumptions concerning the type of sampling strategy that might be logically expected to approximate the stream of behavior. The accumulated evidence raises serious regarding the tenability of some of those logical assumptions. It has been over 10 years, for instance, since questions were first raised about the ability of partial interval sampling to accurately represent proportion of time spent in a behavior. In surveying the literature, however, one finds little evidence that partial interval recording has been abandoned in response to those findings. Similarly, parametric investigations of interval recording have failed to disclose an interval length value that will reliably

and consistently estimate continuous measures. Again, however, the 10 second interval appears to be the accepted standard for behavioral research using partial interval sampling.

For momentary time sampling, inter-observation intervals longer than those typically used will apparently still guarantee accurate results. Although shorter intervals appear to be more accurate, the confounding of interval length and frequency of observation in the majority of investigations precludes firm statements about the ideal interval length. Rather, it is probably more appropriate to conclude that some minimum number of momentary time samples is necessary to ensure accuracy, whether that be achieved through frequent samples drawn from a brief session, or through more intermittent observations spread over a longer period of time. Theoretical accounts have suggested that the "ideal" sampling interval for momentary time sampling also will be dependent on some characteristic or characteristics of the observed behavior. It is not yet clear, however, to what extent the complex interactions of behavioral parameters will allow matching of characteristics to sampling intervals in applied settings.

The continued use of 10 second intervals in momentary time sampling represents not so much a problem of accuracy -- 10 second intervals have been shown to be extremely accurate in comparison to a continuous measure -- as a loss of observational economy. Employing longer intervals between momentary time samples would increase the amount of information researchers could gather per unit time. One of the more encouraging trends in behavioral assessment has been the development of observational systems that focus on both the person and the environment (Greenwood, Delquadri, Stanley, Terry, & Hall, 1985; Saudargas & Lentz, 1986). A longer inter-observation interval could allow greater attention to situational variables between observations of student or client behavior. A shift in sampling strategy might yield benefits for practice as well. The current observational paradigm -- 10 second contiguous intervals during a 30 minute observation period -- requires the presence of an outside observer, usually a

research assistant or psychologist. A more intermittent sampling strategy would be better suited to the schedules of practitioners, facilitating the use of behavioral observation to a greater extent by those practitioners. Finally, there is accumulating evidence that longer observations spaced intermittently over longer sessions would provide a more valid sample of behavior than the contiguous brief sessions currently widely employed.

The issue of representativeness of the behavioral sample remains one of the more unexplored regions of observational assessment. Although early investigations using observational methodology were especially sensitive to issues of sample validity, recent research has tended to rely on relatively brief observation sessions scheduled at a fixed time. It could be argued that the intensive focus on a fixed period for a limited duration has allowed behavioral researchers to gain more precise control over sources of variability in the natural environment (cf. Johnston & Pennypacker, 1980; Sidman, 1960), allowing the methodological refinements that have become something of a hallmark of behavioral assessment (Johnson & Bolstad, 1973; Jones, Reid, & Patterson, 1975). Inferences from a limited temporal sample to some broader universe of an individual's behavior are justified only to the extent that the temporal distributions of behavior in the sample and the universe are roughly equivalent. However, there are no extant data that would support such an assumption. Rather, the available data appear to suggest that the temporal patterning of many behaviors may not be constant over time, and that the current modal observation values--10 to 15 second observation intervals massed in a 30 minute observation period--risk considerable error in estimating behavior occurring over longer periods of time. Given increasing concerns over the generalization of behavioral treatment (Stokes & Osnes, 1986), it would seem appropriate that increased attention be paid to the generalizability of data collected through limited duration behavioral observation.

Implications

An extensive body of time sampling investigations has yielded some fairly

consistent conclusions regarding effective measures and temporal parameters for sampling. While there are fewer resources to draw upon for information about the representativeness of observational sampling, these too suggest a consistent picture. Both literatures provide a rather clear basis for recommendations for the development and use of observational systems for practice or research.

First, momentary time sampling should be used for estimating duration of responding, while event recording should be used for estimating frequency. The superiority of momentary time sampling for estimates of duration has been shown in computer simulations (Green & Alverson, 1978; Harrop & Daniels, 1986), in applied settings (Murphy & Goodall, 1980; Tyler, 1979), and even modeled mathematically (Ary, 1984). There is simply no justification, short of ignorance of the literature, for continued use of partial interval recording to estimate percent of time responding. Given its lack of sensitivity to discrete or brief duration events, however, momentary time sampling appears to be a poor choice for the collection of frequency data. When such data are required, frequency recording should suffice. Combined systems using both momentary time sampling and modified interval recording (Powell, 1984b; Saudargas & Lentz, 1986) appear to hold some promise for observing both duration-based and frequency-based behaviors.

Second, the choice of sampling method should be justified, especially when that choice seems inconsistent with robust experimental findings. It obviously would be unacceptable for a current research report or observation manual to ignore 10-year old findings in the areas of calculation of measures of inter-observer agreement (Hartmann, 1977), reactivity (Mercatoris & Craighead, 1974), or observer drift (DeMaster, Reid, & Twentyman, 1977). Yet the continued use of partial interval recording to estimate duration of response contradicts empirical findings that are equally longstanding (Green & Alverson, 1978; Milar & Hawkins, 1976; Powell et al., 1975, 1977) and highly robust (Dixon, 1981; Dunbar, 1976; Harrop & Daniels, 1986; Kraemer, 1979; Lentz, 1982; Powell, 1984a; Simpson & Simpson, 1977; Tyler, 1979). A fairly detailed rationale would seem to be required of an observational system,

intended for either applied settings or research, that fails to reflect the findings of such a consistent data base.

A related recommendation might be that the choice of sampling interval and session duration be based upon established research findings, not observational tradition. The time-honored tradition of the 10 second interval appears to have little empirical grounding. For partial interval recording, it appears to be too long to guarantee accuracy; for momentary time sampling, it is probably shorter than necessary for efficient observation. Choice of session duration seems similarly to be dictated more by questions of observational convenience than technical adequacy. In order to ensure generalizability, investigators should demonstrate, either experimentally or by reference to the literature, that observation is sufficiently frequent within a sufficiently representative observation period, rather than simply to rely on observational convention.

Finally, increased attention must be paid to the representativeness of the observational sample. Failure to attend to questions of sample representativeness poses a serious threat to the generalizability of behavioral research. Stability of behavior over time and situations represents an important universe of generalization (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) that cannot be ignored. Research reports should indicate to what population of behavior the results can be expected to generalize, and provide justification for that assertion. Failure to address questions of sample representativeness represents a serious threat to the external validity of research results. For practitioners, attention to representativeness of the observational sample may be even more critical. Attention to behavior in only one stimulus situation probably limits the ecological validity of both assessment and intervention (Foster & Cone, 1980; Rogers-Warren & Warren, 1977).

Generalizability theory (Cronbach et al., 1972; Jones et al., 1975; McGaw, Wardrop, & Bunda, 1972), wherein a number of components that contribute to reliability and validity are considered simultaneously to assess their relative contribution to the variance of obtained scores, could provide a powerful tool for

identifying facets of measurement contributing to the accuracy and representativeness of time sampling. Generalizability studies have shown some promise for specifying the number of occasions or raters necessary to obtain reliable ratings of teaching behavior (Erlich & Borich, 1979; Shavelson & Dempsey-Atwood, 1976; Smith, Waller, & Waller, 1982). Similar models could be applied to investigate the generalizability of observational measures typically used in single case research. Such an approach is not necessarily antithetical to the idiographic focus of behavioral assessment. Rather, generalizability studies provide a useful description of the sources of variability that must be taken into account in the observation of behavior (Cone, 1979; Cone & Foster, 1982).

Ultimately, both the accuracy and representativeness of time sampling resolve into questions of generalizability. Cronbach et al. (1972) note that the most appropriate measurement question is not whether an obtained score is accurate with respect to a "true score", but rather whether the obtained data is generalizable to the universe of behavior or situations of interest to the investigator. The accuracy of a given sampling strategy is primarily a question of temporal generalization. Will behavioral data obtained by sampling some portion of a time interval generalize to the universe of behavior represented by the remainder of that interval? The issue of representativeness is more a matter of situational generalizability. To what extent will data from one situation generalize to behavior occurring under alternate conditions? How many situations must be observed to acquire data that are generalizable to the universe of behavior in which the experimenter is interested? Ensuring generalizability of observation along these and other dimensions requires careful thought and planning in the development and implementation of an observational system, and probably further parametric investigation.

Still, the literature investigating the adequacy of time sampling strategies has made considerable progress in identifying parameters that influence the accuracy of observational samples. Given the complexity of human behavior, and the equally perplexing complexity of controlling contingencies, notable successes in generating

robust findings in this area might be regarded as more remarkable than the failures. Certainly further advances in specifying temporal parameters critical to generalizability of observational data can be expected. Perhaps the more important challenge, however, is for observational application to catch up with observational theory.

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. Behaviour, 49, 227-267.

Arrington, R. E. (1939). Time sampling studies of child behavior. Psychological Monographs, 51 (2).

Arrington, R. E. (1943). Time sampling in studies of social behaviors: A critical review of techniques and results with research suggestions. Psychological Bulletin, 40, 81-124.

Ary, D. (1984). Mathematical explanation of error in duration recording using partial interval, whole interval, and momentary time-sampling. Behavioral Assessment, 6, 221-228.

Ary, D. & Suen, H. K. (1983). The use of momentary time sampling to assess both frequency and duration of behavior. Journal of Behavioral Assessment, 5, 143-150.

Baer, D. M. (1986). In application, frequency is not the only estimate of the probability of behavior units. In T. Thompson & M. D. Zeiler (Eds.), Analysis and integration of behavioral units (pp. 117-136). Hillsdale, N. J.: Lawrence Erlbaum.

Baer, D. M. & Fowler, S. A. (1984). How should we measure the potential of self-control procedures for generalized educational outcomes? In W. L. Heward, T. E. Heron, D. S. Hill, & J. Trap-Porter (Eds.), Focus on behavior analysis in education. Columbus, OH: Charles E. Merrill.

Bass, R. G. (1987). Computer-assisted observer training. Journal of Applied Behavior Analysis, 20, 83-88.

Baum, C. G., Forehand, R., & Zegiob, L. E. (1979). A review of observer reactivity in adult-child interactions. Journal of Behavioral Assessment, 1, 167-178.

Bem, D. J. (1972), Constructing cross-situational consistencies in behavior: Some thoughts on Alker's critique of Mischel. Journal of Personality, 40, 17-26.

Berne, E. V. (1930). An experimental investigation of social behavior patterns in young children. University of Iowa Studies in Child Welfare, (4, No. 3).

Bijou, S. W., Peterson, R. F. , & Ault, M. H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. Journal of Applied Behavior Analysis, 1, 175-191.

Bindra, D. & Blond, J. (1958). A time-sample method for measuring general activity and its components. Canadian Journal of Psychology, 12, 74-76.

Bott, E. A., Blatz, W. E., Chant, N., & Bott, H. M. (1928). Observation and training of fundamental habits in young children. Genetic Psychology Monographs, 4, 5-161.

Brulle, A. R. & Repp, A. C. (1984). An investigation of the accuracy of momentary time sampling procedures with time series data. British Journal of Psychology, 75, 481-485.

Bushell, D., Wrobel, P. A., & Michealis, M. L. (1968). Applying "group" contingencies to the classroom study behavior of preschool children. Journal of Applied Behavior Analysis, 1, 55-61.

Cochran, W.G. (1977). Sampling techniques (3rd ed.). New York: Wiley.

Cone, J.D. (1979). Confounded comparisons in triple response mode assessment research. Behavioral Assessment, 1, 85-95.

Cone, J.D. & Foster, S.L. (1982). Direct observation in clinical psychology. In P.C. Kendall & J.N. Butcher (Eds.), Handbook of research methods in clinical psychology. New York: Wiley.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurement: Generalizability of scores and profiles. New York: Wiley.

DeMaster, B., Reid, J., & Twentyman, C. (1977). The effects of different amounts of feedback on observer's reliability. Behavior Therapy, 8, 317-329.

Dixon, J. W. (1981). A comparison of thirteen time sampling systems with continuous real time measures. Dissertation Abstracts International, 42, 2050-B.

Drabman, R. & Spitalnik, R. (1973). Social isolation as a punishment procedure: A controlled study. Journal of Experimental Child Psychology, 16, 236-249.

Dunbar, R. (1976). Some aspects of research design and their implications in the observational study of behavior. Behaviour, 58, 79-98.

Erlich, O. & Borich, G. (1979). Occurence and generalizability of scores on a classroom interaction instrument. Journal of Educational Measurement, 16, 11-18.

Erlich, O. & Shavelson, R.J. (1978). The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualization problem, or both? Journal of Educational Measurement, 15, 77-89.

Estep, D. Q., Johnston, M. E., & Gordon, T. P. (1981). The effectiveness of sampling methods in detecting copulatory behavior in *Macaca arctoides*. American Journal of Primatology, 1, 453-455.

Foster, S. L. & Cone, J. D. (1980). Current issues in direct observation. Behavioral Assessment, 2, 313-338.

Foster, S. L. & Cone, J. D. (1986). Design and use of direct observation procedures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), Handbook of Behavioral Assessment (2nd ed.) (pp. 253-324). New York: Wiley.

Goldfried, M. R. (1977). Behavioral assessment in perspective. In J. D. Cone & R. P. Hawkins (Eds.), Behavioral Assessment: New Directions in Clinical Psychology (pp. 3-22). New York: Brunner/Mazel.

Goldfried, M. R. & Kent, R. N. (1972),. Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. Psychological Bulletin, 77, 409-420.

Goodenough, F. L. (1928). Measuring behavior traits by means of repeated short samples. Journal of Juvenile Research, 12, 230-235.

Goodenough, F. L. (1930). Inter-relationships in the behavior of young children. Child Development, 1, 29-47.

Green, S. B. & Alverson, L. G. (1978). A comparison of measures for long-duration behaviors. Journal of Applied Behavior Analysis, 11, 530.

Green, S. B., McCoy, J. F., Burns, K. P., & Smith, A. C. (1982). Accuracy of observational data with whole interval, partial interval, and momentary time-sampling recording techniques. Journal of Behavioral Assessment, 4, 103-118.

Greenwood, C. R., Delquadri, J. C., Stanley, S. O., Terry, B. & Hall, R. V. (1985). Assessment of eco-behavioral interaction in school settings. Behavioral Assessment, 7, 331-347.

Griffin, B. & Adams, R. (1983). A parametric model for estimating prevalence, incidence, and mean bout duration from point sampling. American Journal of Primatology, 4, 261-271.

Hall, R. V., Lund, D., & Jackson, D. (1968). Effects of teacher attention on study behavior. Journal of Applied Behavior Analysis, 1, 1-12.

Hallahan, D. P., Marshall, K. J., & Lloyd, J. W. (1981). Self-recording during group instruction: Effects on attention to task. Learning Disability Quarterly, 4, 407-413.

Harrop, A. & Daniels, M. (1985). Momentary time sampling with time series data: A commentary on the paper by Brulle & Repp. British Journal of Psychology, 76, 533-537.

Harrop, A. & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. Journal of Applied Behavior Analysis, 19, 73-77.

Hartmann, D.P. (1977). Considerations in the use of interobserver reliability estimates. Journal of Applied Behavior Analysis, 10, 103-116.

Haynes, S. N. (1978). Principles of behavioral assessment. New York: Gardner.

Hoge, R. D. (1985). The validity of direct observation measures of pupil classroom behavior.
Review of Educational Research, 55, 469-484.

Johnson, S. M. & Bolstad, O. D. (1973). Methodological issues in naturalistic observation: Some problems for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology, concepts, and practice (pp. 7-68). Champaign, IL: Research Press.

Johnston, J. M. & Pennypacker, H. S. (1980). Strategies and tactics of human behavioral research. Hillsdale, NJ: Lawrence Erlbaum.

Jones, R. R., Reid, J. B., & Patterson, G. R. (1975). Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment (Vol. 3). San Francisco: Jossey-Bass.

Karweit, N. & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition. Journal of Educational Psychology, 74, 844-851.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABC's of reliability. Journal of Applied Behavior Analysis, 10, 141-150.

Kazdin, A. E. (1979). Situation specificity: The two-edged sword of behavioral assessment. Behavioral Assessment, 1, 57-75.

Kazdin, A. E. (1982). Single-case research designs: Methods for clinical and applied settings.
New York: Oxford University.

Kelly, M. B. (1977). A review of the observational data-collection and reliability procedures reported in *The Journal of Applied Behavior Analysis*. Journal of Applied Behavior Analysis, 10, 97-101.

Kent, R. N., O' Leary, K. D., Diament, C., & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. Journal of Consulting and Clinical Psychology, 42, 774-780.

Kraemer, H. (1979). One-zero sampling in the study of primate behavior. Primates, 20, 237-244.

Kubany, E. S. & Sloggett, B.B. (1973). Coding procedures for teachers. Journal of Applied Behavior Analysis, 6, 339-344.

Leach, D. J., & Dolan, N. G. (1985). Helping teachers increase student academic engagement rate: The evaluation of a minimal feedback procedure. Behavior Modification, 9, 55-71.

Leger, D. (1977). An empirical evaluation of instantaneous and one-zero sampling of chimpanzee behavior. Primates, 20, 387-393.

Lentz, F. E. (1982). An empirical examination of the utility of partial interval and momentary time sampling as measurements of behavior. Dissertation Abstracts International, 43, 545-B.

Lomax, R.G. & Cooley, W.W. (1979, April). The student achievement-instructional time relationship. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Mansell, J. (1985). Time sampling and measurement error: The effect of interval length and sampling pattern. Journal of Behaviour Therapy and Experimental Psychiatry, 16, 245-251.

Mattos, R. L. (1971). Some relevant dimensions of interval recording. Academic Therapy, 6, 235-244.

McDowell, E. (1973). Comparison of time-sampling and continuous recording techniques for observing developmental changes in caretaker and infant behaviors. The Journal of Genetic Psychology, 123, 99-105.

McGaw, B., Wardrop, J.L., & Bunda, M.A. (1972). Classroom observation schemes--where are the errors? American Educational Research Journal, 9, 13-27.

Mercatoris, M. & Craighead, W.E. (1974). The effects of nonparticipant observation on teacher and pupil classroom behavior. Journal of Educational Psychology, 66, 512-519.

Milar, C. R. & Hawkins, R. P. (1976). Distorted results from the use of interval scoring procedures. In T. A. Brigham, R. Hawkins, J. W. Scott, & T. F. McLaughlin (Eds.), Behavior analysis in education: Self-control and reading (pp. 261-273). Dubuque, IA: Kendall/Hunt.

Mischel, W. (1968). Personality and assessment. New York: Wiley.

Mischel, W. (1973). Toward a cognitive social learning conceptualization of personality. Psychological Review, 4, 252-283.

Murphy, G. & Goodall, E. (1980). Measurement error in direct observation: A comparison of common recording methods. Behaviour Research and Therapy, 18, 147-150.

Nelson, R. O. (1985). Behavioral assessment in the school setting. In T. R. Kratochwill (Ed.), Advances in school psychology (pp. 45-87). Hillsdale, NJ: Lawrence Erlbaum.

O' Leary, K. D., Drabman, R. S., & Kass, R. E. (1973). Maintenance of appropriate behavior in a token program. Journal of Abnormal Child Psychology, 1, 127-138.

O' Leary, K. D., Kaufman, K. F., Kass, R. E., & Drabman, R. (1970). The effects of loud and soft reprimands on the behavior of disruptive students. Exceptional Children, 37, 145-155.

Olson, W. C. (1929). The measurement of nervous habits in normal children (Monograph No. 3). Minneapolis: University of Minnesota Institute of Child Welfare.

Olson, W. C. (1931). A study of classroom behavior. Journal of Educational Psychology, 22, 449-454.

Parke, R. D. (1979). Interactional designs. In R. B. Cairns (Ed.), The analysis of social interactions: Methods, issues, and illustrations (pp. 15-36). Hillsdale, NJ: Lawrence Erlbaum.

Parten, M. B. (1932). Social participation among preschool children. Journal of Abnormal Social Psychology, 27, 243-269.

Powell, J. (1984a). On the misrepresentation of behavioral realities by a widely practiced direct observation procedure: Partial interval (one-zero) sampling. Behavioral Assessment, 6, 209-219.

Powell, J. (1984b). Some empirical justification for a modest proposal regarding data acquisition via intermittent direct observation. Journal of Behavioral Assessment, 6, 71-80.

Powell, J., Martindale, A. & Kulp, S. (1975). An evaluation of time-sample measures of behavior. Journal of Applied Behavior Analysis, 8, 463-469.

Powell, J., Martindale, B. Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. Journal of Applied Behavior Analysis, 10, 325-332.

Powell, J. & Rockinson, R. (1978). On the inability of interval time sampling to reflect frequency of occurrence data. Journal of Applied Behavior Analysis, 11, 531-532.

Ragland, E. U., Kerr, M. M., & Strain, P. S. (1981). Social play of withdrawn children: A study of teacher-mediated peer feedback. Behavior Modification, 5, 347-359.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval, and time-sampling methods of data collection. Journal of Applied Behavior Analysis, 9, 501-508.

Rhine, R. J. & Ender, P. B. (1983). Comparability of methods used in the sampling of primate behavior. American Journal of Primatology, 5, 1-15.

Rhine, R. & Flanigon, M. (1978). An empirical comparison of one-zero, focal-animal, and instantaneous methods of sampling spontaneous primate social behavior. Primates, 19, 353-361.

Rhine, R. & Linville, A. (1980). Properties of one-zero scores in observational studies of primate social behavior: The effect of assumptions on empirical analysis. Primates, 21, 111-122.

Risley, T. R. (1971). Spontaneous language in the preschool environment. In J. Stanley (Ed.), Research on curriculums for preschools. Baltimore: Johns Hopkins.

Rogers-Warren, A. & Warren, S.F. (1977). Ecological perspectives in behavior analysis. Baltimore: University Park Press.

Rojahn, J. & Kanoy, R.C. (1985). Toward an empirically based parameter selection for time-sampling observation systems. Journal of Psychopathology and Behavioral Assessment, 7, 99-120.

Rowley, G. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brwon formula. Journal of Educational Measurement, 15, 165-180.

Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. Journal of Applied Behavior Analysis, 13, 493-500.

Saudargas, R.A. & Lentz, F.E. (1986). Estimating percent of time and rate via direct observation: A suggested observational procedure and format. School Psychology Review, 15, 36-48.

Schachar, R., Sandberg, S., & Rutter, M. (1986). Agreement between teachers' ratings and observations of hyperactivity, inattentiveness, and defiance. Journal of Abnormal Child Psychology, 14, 331-345.

Shavelson, R. & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. Review of Educational Research, 46, 553-611.

Sidman, M. (1960). Tactics of scientific research: Evaluating experimental data in psychology. New York: Basic Books.

Simpson, M. & Simpson, A. (1977). One-zero and scan methods for sampling behavior. Animal Behaviour, 25, 726-731.

Skiba, R. J. (1987). Accuracy and representativeness of behavioral observation as a function of sampling strategy and type of behavior. Unpublished doctoral dissertation, University of Minnesota.

Skiba, R. J. & Casey, A. (1985). Interventions for behavior disordered students: A quantitative review and methodological critique. Behavioral Disorders, 10, 239-252.

Skiba, R. J., Casey, A., & Center, B. A. (1985/1986). Nonaversive procedures in the treatment of classroom behavior disorders. The Journal of Special Education, 19, 459-482.

Skiba, R. J. & O'Sullivan, P. J. (1987). Implications of the relationship between observational and rating scale data for classroom assessment. In S. Braaten, R.B. Rutherford, II, & J. Maag (Eds.), Programming for adolescents with behavior disorders, Vol. 3 (pp.5-15). Reston, VA: Council for Children with Behavioral Disorders.

Smith, P. L., Waller, M. I, & Waller, S. P. (1982). Generalizable observation of the teaching process. Educational and Psychological Measurement, 42, 467-478.

Stokes, T. F. & Osnes, P. G. (1986). Programming the generalizatio nof children's social behavior. In P.S. Strain, M.J. Guralinick, & H.M. Walker (Eds.), Children's social behavior: Development, assessment, and modification (pp. 407-443). Orlando, FLA: Academic Press.

Suen, H. K. (1986). On the utility of a *post hoc* correction procedure for one-zero sampling duration estimates. Primates, 27, 237-244.

Suen, H. K. & Ary, D. (1984). Variables influencing one-zero and instantaneous time-sampling outcomes. Primates, 2, 89-94.

Suen, H. K. & Ary, D. (1986). Poisson cumulative probabilities and systematic errors in single-subject and multiple-subject time sampling. Behavioral Assessment, 8, 155-169.

Test, D. W. & Heward, W. L. (1984). Accuracy of momentary time-sampling: Comparison of fixed- and variable-interval observation schedules. In W. L. Heward, T. E. Heron, D. S. Hill, & J. Trap-Porter (Eds.), Focus on behavior analysis in education (pp. 177-194). Columbus, OH: Charles E. Merrill.

Thiemann, S. & Kraemer, H. C. (1984). Sources of behavioral variance: Implications for sample size decisions. American Journal of Primatology, 7, 367-375.

Thomson, C., Holmberg, M., & Baer, D. M. (1974). A brief report on a comparison of time-sampling procedures. Journal of Applied Behavior Analysis, 7, 623-626.

Tindal, G. & Parker, R. (1987). Direct observation in special education classrooms: Concurrent use of two instruments and their validation. The Journal of Special Education, 21, 43-58.

Tobin, K.G. & Capie, W. (1981, April). Measuring pupil engagement. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Tyler, S. (1979). Time-sampling: A matter of convention. Animal Behaviour, 27, 801-810.

Walker, H. M. & Lev, J. (1953). Statistical inference. New York: Holt, Rinehart, & Winston.

Wasik, B. H. (1984). Clinical applications of direct behavioral observation. Advances in Clinical Child Psychology, 7, 153-193.

Wildman, B. G. & Erickson, M. T. (1977). Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkins (Eds.), Behavioral Assessment: New Directions in Clinical Psychology (pp. 255-273). New York: Brunner/Mazel.

Witt, J. C. & Adams, R. M. (1980). Direct and observed reinforcement in the classroom: The interaction between information and reinforcement for socially approved and disapproved behaviors. Behavior Modification, 4, 321-326.

FOOTNOTES

The author is now at Indiana University, Urbana-Champaign. This paper was originally part of the author's dissertation. Appreciation is extended to James Ysseldyke, Frank Wood, and Maynard Reynolds for their comments on an earlier draft of this manuscript.

[1] There has been some confusion regarding the appropriate terminology for time sampling. Some have distinguished between interval recording as the observation of occurence or non-occurence of a behavior during an entire interval of time, and time sampling as the observation of occurence or non-occurence at a specific point in time (Repp, Roberts, Slack, Repp, & Berkler, 1976). Others have designated all such methods under the generic label of time-sampling (Powell, Martindale & Kulp, 1975). As the labels and definitions suggested by Powell et al. (1975)--partial interval sampling, whole interval sampling, and momentary time sampling--have become widely accepted in the behavioral assessment literature (Foster & Cone, 1986), the convention of referring to all three methods as time sampling will also be adopted here. For convenience, however, both whole and partial interval sampling will be referred to as interval recording or interval sampling.

[2] Although the terminological conventions suggested by Powell et al.(1975) have been widely adopted, a fourth time sampling methodology, predominant activity sampling, has been identified since that typology was introduced. The behavior counted as an occurrence in predominant activity sampling is that behavior that occupies the majority of time in the defined interval. While predominant activity sampling appears to have some conceptual advantages in the observation of more than one target behavior, there have been few experimental (Sanson-Fisher, Poole, & Dunn, 1980; Tyler, 1979), and no intervention-oriented investigations of the procedure. Given the paucity of data on this method, the remainder of the review will focus on the three more widely documented methods of time-sampling.

IAP PUBLICATIONS

Instructional Alternatives Project
350 Elliott Hall
University of Minnesota
75 East River Road
Minneapolis, MN 55455

Research Reports

No. 1 Time allocated to instruction of mentally retarded, learning disabled, emotionally disturbed, and nonhandicapped elementary students by J. E. Ysseldyke, M. L. Thurlow, S. L. Christenson, & J. Weiss (March, 1987).

No. 2 Instructional tasks used by mentally retarded, learning disabled, emotionally disturbed, and nonhandicapped elementary students by J. E. Ysseldyke, S. L. Christenson, M. L. Thurlow, & D. Bakewell (June, 1987).

No. 3 Instructional grouping arrangements used with mentally retarded, learning disabled, emotionally disturbed, and nonhandicapped elementary students by J. E. Ysseldyke, M. L. Thurlow, S. L. Christenson, & R. McVicar (July, 1987).

No. 4 Academic engagement and active responding of mentally retarded, learning disabled, emotionally disturbed and nonhandicapped elementary students by J. E. Ysseldyke, S. L. Christenson, M. L. Thurlow, & R. Skiba (July, 1987).

No. 5 The qualitative nature of instruction for mentally retarded, learning disabled, and emotionally disturbed elementary students in special education by J. E. Ysseldyke, S. L. Christenson, & M. L. Thurlow (July, 1987).

No. 6 State guidelines for student-teacher ratios for mildly handicapped children by M. L. Thurlow, J. E. Ysseldyke, & J. W. Wotruba (July, 1987).

No. 7 Student-teacher ratios for mildly handicapped children in special education settings by J. E. Ysseldyke, M. L. Thurlow, & J. W. Wotruba (November, 1987).

No. 8 Regular education teachers' perceptions of instructional arrangements for students with mild handicaps by J. E. Ysseldyke, M. L. Thurlow, J. W. Wotruba, & P. A. Nania (January, 1988).

No. 9 Differences in the qualitative nature of instruction for LD and EBD students in regular and special education settings by J. E. Ysseldyke, S. L. Christenson, & M. L. Thurlow (January, 1988).

No. 10 Alternate explanations for learning disabled, emotionally disturbed, and educable mentally retarded students' reading achievement by J. E. Ysseldyke, D. Bakewell, S. L. Christenson, P. Muyskens, J. G. Shriner, M. Cleary, & J. Weiss (July, 1988).

No. 11 Alternate explanations for learning disabled, emotionally disturbed, and educable mentally retarded students' math achievement by J. E. Ysseldyke, M. Cleary, S. L. Christenson, P. Muyskens, J. G. Shriner, D. Bakewell, & J. Weiss (August, 1988).

No. 12 Student and instructional outcomes under varying student-teacher ratios in special education by M. L. Thurlow, J. E. Ysseldyke, & J. W. Wotruba (August, 1988).

No. 13 Teacher stress and student achievement for mildly handicapped students by D. Bakewell, S. R. McConnell, J. E. Ysseldyke, & S. L. Christenson (August, 1988).

IAP PUBLICATIONS

Page Two

No. 14 A case study analysis of factors related to effective student-teacher ratios by J. E. Ysseldyke, M. L. Thurlow, J. G. Shriner, & C. S. Propsom (August, 1988).

No. 15 Written language: The instructional experience of mildly handicapped and nonhandicapped elementary students by R. McVicar, S. L. Christenson, M. L. Thurlow, & J. E. Ysseldyke (August, 1988).

No. 16 Social validity of different student-teacher ratios by M. L. Thurlow, J. E. Ysseldyke, & C. Yeh (August 1988).

No. 17 Home environments of mildly handicapped and nonhandicapped elementary students by S. L. Christenson, J. E. Ysseldyke, & M. Cleary (September, 1988).

No. 18 Volunteer tutors as a reading intervention for sudents with reading difficulties by J. Weiss, M. L. Thurlow, S. L. Christenson, & J. E. Ysseldyke (October, 1988).

Monographs

No. 1 Instructional environment scale: Scale development and training procedures by J. E. Ysseldyke, S. L. Christenson, R. McVicar, D. Bakewell, & M. L. Thurlow (December, 1986).

No. 2 Instructional psychology and models of school learning: Implications for effective instruction of handicapped students by S. L. Christenson, J. E. Ysseldyke, & M. L. Thurlow (May, 1987).

No. 3 School effectiveness: Implications for effective instruction of handicapped students by M. L. Thurlow, S. L. Christenson, & J. E. Ysseldyke (May, 1987).

No. 4 Instructional effectiveness: Implications for effective instruction of handicapped students by S. L. Christenson, M. L. Thurlow, & J. E. Ysseldyke (May, 1987).

No. 5 Teacher effectiveness and teacher decision making: Implications for effective instruction of handicapped students by J. E. Ysseldyke, M. L. Thurlow, & S. L. Christenson (May, 1987).

No. 6 Student cognitions: Implications for effective instruction of handicapped students by M. L. Thurlow, J. E. Ysseldyke, & S. L. Christenson (May, 1987).

No. 7 Instructional factors that influence student achievement: An integrative review by J. E. Ysseldyke, S. L. Christenson, & M. L. Thurlow (September, 1987).

No. 8 Adults in the classroom: Effects on special education instruction by A. E. Dear, M. L. Thurlow, & J. E. Ysseldyke (September, 1987).

No. 9 Student-teacher ratios and their relationship to instruction and achievement for mildly handicapped students. Final Report by J. E. Ysseldyke (August, 1988).

No. 10 Temporal parameters in the sampling of behavior: The accuracy and generalizability of observation by R. J. Skiba (April, 1989).